CLAIMS

What is claimed is:

1. A method for reducing lexical ambiguity in an input stream, comprising:

breaking the input stream into at least two tokens;

creating a connection graph using the at least two tokens, the connection graph comprising a plurality of paths;

assigning a cost to each of the plurality of paths;

defining at least one best path based upon a corresponding cost to generate an output graph; and

providing the output graph to a syntactic analysis module to reduce lexical ambiguity.

2. The method of claim 1 wherein a number of the at least one best path is either predefined or determined programmatically.

3. The method of claim 1 wherein creating a connection graph using the at least two tokens comprises:

compiling lexical grammar rules to generate lexical functions, the lexical grammar rules being written in a grammar programming language;

creating a plurality of segments from the at least two tokens based upon lexical information and the lexical functions, and

defining the plurality of paths using the plurality of segments.

4. The method of claim 1 wherein creating a connection graph using the at least two tokens comprises assigning at least one part of speech tag to at least one of the at least two tokens using lexical information.

5. The method of claim 1 wherein creating a connection graph using the at least two tokens comprises recognizing a multiword expression in the input stream using multiword information.

6. The method of claim 1 wherein the connection graph comprises a set of nodes and a set of arcs.

7. The method of claim 6 wherein each of the plurality of paths comprises a combination of nodes and arcs.

8. The method of claim 1 wherein the cost comprises lexical cost, unigram cost, bigram cost and connector cost.

9. A method for providing segmentation of an input stream having at least two tokens, comprising:

creating a plurality of segments from the at least two tokens based upon lexical information and lexical functions; and

generating a connection graph using the plurality of segments.

10. The method of claim 9 further comprising compiling lexical grammar rules to generate the lexical functions, the lexical grammar rules being written in a grammar programming language.

11. The method of claim 10 wherein the lexical grammar rules define connectivity relation of tokens.

12. The method of claim 9 further comprising assigning at least one part of speech tag to at least one segment using a lexical dictionary.

13. The method of claim 12 further comprising:

defining a plurality of paths in the connection graph based upon part of speech tags and the segments;

assigning a cost to each of the plurality of paths; and

determining at least one best path based upon a corresponding cost to generate an output graph.

14. An apparatus for reducing lexical ambiguity in an input stream, comprising:

means for breaking the input stream into at least two tokens;

means for creating a connection graph using the at least two tokens, the connection graph comprising a plurality of paths;

means for assigning a cost to each of the plurality of paths;

means for defining at least one best path based upon a corresponding cost to generate an output graph; and

means for providing the output graph to a syntactic analysis module to reduce lexical ambiguity.

15. The apparatus of claim 14 wherein a number of the at least one best path is either predefined or determined programmatically.

16. The apparatus of claim 14 further comprising:

means for compiling lexical grammar rules to generate lexical functions, the lexical grammar rules being written in a grammar programming language;

means for creating a plurality of segments from the at least two tokens based upon lexical information and the lexical functions, and

means for defining the plurality of paths using the plurality of segments.

17. The apparatus of claim 14 further comprising means for assigning at least one part of speech tag to at least one of the at least two tokens using lexical information.

18. The apparatus of claim 14 further comprising means for recognizing a multiword expression in the input stream using multiword information.

19. The apparatus of claim 14 wherein the connection graph comprises a set of nodes and a set of arcs.

20. The apparatus claim 19 wherein each of the plurality of paths comprises a combination of nodes and arcs.

21. The apparatus of claim 14 wherein the cost comprises lexical cost, unigram cost, bigram cost and connector cost.

22. An apparatus for providing segmentation of an input stream having at least two tokens, comprising:

means for creating a plurality of segments from the at least two tokens based upon lexical information and lexical functions; and

means for generating a connection graph using the plurality of segments.

23. The apparatus of claim 22 further comprising means for compiling lexical grammar rules to generate the lexical functions, the lexical grammar rules being written in a grammar programming language.

24. The apparatus of claim 23 wherein the lexical grammar rules define connectivity relation of tokens.

25. The apparatus of claim 22 further comprising means for assigning at least one part of speech tag to at least one segment using a lexical dictionary.

26. The apparatus of claim 25 further comprising:

means for defining a plurality of paths in the connection graph based upon part of speech tags and the segments;

means for assigning a cost to each of the plurality of paths; and

means for determining at least one best path based upon a corresponding cost to generate an output graph.

27. An apparatus for reducing lexical ambiguity in an input stream, comprising:

a tokenizer for breaking the input stream into at least two tokens;

a token connector for creating a connection graph using the at least two tokens, the connection graph comprising a plurality of paths;

a cost assignor for assigning a cost to each of the plurality of paths;

a path calculator for defining at least one best path based upon a corresponding cost to generate an output graph; and

a graph provider for providing the output graph to a syntactic analysis module to reduce lexical ambiguity.

28. The apparatus of claim 27 wherein a number of the at least one best path is either predefined or determined programmatically.

29. The apparatus of claim 27 wherein the token connector comprises:

a grammar programming language (GPL) compiler for compiling lexical grammar rules to generate lexical functions, the lexical grammar rules being written in a general programming language;

a segmentation engine for creating a plurality of segments from the at least two tokens based upon lexical information and the lexical functions, and

a path designator for defining the plurality of paths using the plurality of segments.

30. The apparatus of claim 27 wherein the token connector comprises a part of speech tagger for assigning at least one part of speech tag to at least one of the at least two tokens using lexical information.

31. The apparatus of claim 27 wherein the token connector comprises a multiword recognizer for recognizing a multiword expression in the input stream using multiword information.

32. The apparatus of claim 27 wherein the connection graph comprises a set of nodes and a set of arcs.

33. The apparatus of claim 32 wherein each of the plurality of paths comprises a combination of nodes and arcs.

34. The apparatus of claim 27 wherein the cost comprises lexical cost, unigram cost, bigram cost and connector cost.

35. An apparatus for providing segmentation of an input stream having at least two tokens, comprising:

a segmentation engine for creating a plurality of segments from the at least two tokens based upon lexical information and lexical functions; and

a graph generator for generating a connection graph using the plurality of segments.

36. The apparatus of claim 35 further comprising a grammar programming language (GPL) compiler for compiling lexical grammar rules to generate the lexical functions, the lexical grammar rules being written in GPL.

37. The apparatus of claim 36 wherein the lexical grammar rules define connectivity relation of tokens.

38. The apparatus of claim 35 further comprising a part of speech tagger for assigning at least one part of speech tag to at least one segment using lexical information.

39. The apparatus of claim 38 further comprising:

a path designator for defining a plurality of paths in the connection graph based upon part of speech tags and the segments;

a cost assignor for assigning a cost to each of the plurality of paths; and

a path calculator for determining at least one best path based upon a corresponding cost to generate an output graph.

40. A system for reducing lexical ambiguity, comprising:

a processor;

an input coupled to the processor, the input capable of receiving an input stream, the processor configured to break the input stream into at least two tokens, create a connection graph comprising a plurality of paths using the at least two tokens, assign a cost to each of the plurality of paths, and define at least one best path based upon a corresponding cost to generate an output graph; and

an output coupled to the processor, the output capable of providing the output graph to a syntactic analysis module to reduce lexical ambiguity.

41. A system for providing segmentation of an input stream, comprising:

a processor;

an input coupled to the processor, the input capable of receiving an input stream having at lest two tokens, the processor configured to create a plurality of segments from the at least two tokens based upon lexical information and lexical functions, and generate a connection graph using the plurality of segments; and

an output coupled to the processor, the output capable of providing segmentation of the input stream.

42. A computer readable medium comprising instructions, which when executed on a processor, perform method for reducing lexical ambiguity in an input stream, comprising:

breaking an input stream into at least two tokens;

creating a connection graph using the at least one token, the connection graph comprising a plurality of paths;

assigning a cost to each of the plurality of paths;

defining at least one best path based upon a corresponding cost to generate an output graph; and

providing the output graph to a syntactic analysis module to reduce lexical ambiguity.

43. The computer readable medium of claim 42 wherein creating a connection graph further comprises providing segmentation of the input stream using lexical information and lexical functions.

44. The computer readable medium of claim 42 wherein creating a connection graph further comprises assigning at least one part of speech tag to at least one of the at least two tokens using lexical information.

45. The computer readable medium of claim 42 wherein creating a connection graph further comprises recognizing a multiword expression in the input stream using lexical information.

46. The computer readable medium of claim 42 wherein a number of the at least one best path is either predefined or determined programmatically.

47. A computer readable medium comprising instructions, which when executed on a processor, perform method for providing segmentation of an input stream having at least two tokens, comprising:

creating a plurality of segments from the at least two tokens based upon lexical information and lexical functions; and

generating a connection graph using the plurality of segments.

48. The computer readable medium of claim 47 further comprising compiling the lexical grammar rules to generate lexical functions, the lexical grammar rules being written in a grammar programming language.

49. A memory for storing data for access by an application program being executed on a data processing system, comprising:

a data structure stored in said memory, said data structure including information resident in a file used by said application program and including:

a plurality of packet structures used for the transmission of data, wherein each packet structure includes

a set of nodes,

a set of arcs connecting at least two of the set of nodes, and

a value data object for each of the set of arcs having a value that represents a corresponding part of speech tag.